Big Data Approach to Traffic Violation Data Analytics.

Prasad Modi, Abolanle Sule, Swetha Rao, Advisor: Prof. Jeongkyu Lee.
Department of Computer Science, University of Bridgeport, 06604 CT, United States.

Abstract

This paper presents an in-depth analysis of traffic violations patterns using Big Data Analytic methods, aimed at improving monitoring patrol, government planning and decision making. The analysis was carried out using data available from over a period of four years in all US states. It describes certain problem definitions, processing and analyzing unstructured data using Hadoop Systems tools, data visualization and evaluating the results. In addition we described the MapReduce programming model in relation to our data analysis process and provided its implementation details. We implemented other Hadoop Eco System tools such as Hive and Spark that were used for comparison and evaluation. Finally, this paper presents the data visualization concept use to further present data in graphical forms for better pattern identification.

Problem Definition & Data Set

We are interested if finding various traffic violations patterns, this patterns will assist intensive monitoring, planning and decision making: Some of the queries we will be implementing includes the following:

- Traffic violations based on time of violation, date of violation and gender, violations having a particular pattern related to time when the violation is high.
- Traffic violations based on the gender. Whether it is myth or reality about female drivers are not as good as male behind the wheels along with violation type ?
- Traffic violations which resulted into accidents, Number of accidents by year.

Data Set

We are analysing traffic violation records, which we have gotten from government released data set. Link for the data set is

https://catalog.data.gov/dataset/traffic-violations-56dda

Design And Implementation

We present a simple modelling of the MapReduce programming and most importantly how it relates to our analysis. Its enhanced with sorting, grouping, reduction capabilities and scales easily. Its structured for parallel computing, high failure tolerances and assists in hiding the messy details from users. It has three main components the mapper, shuffle-sort and the reducer. Conceptually, it leaves two functions to be implemented by the user.[4] Mapper function takes a key,value pair and transform to a list of key,value pair. The shuffle- sort is performed after all map jobs are completed, this phase sorts and consolidate the intermediate data and send to the reducer. Reducer function transforms the intermediate data(key,iterable value pair) and writes the final output data (key,value) pair.

Implementation Chart for the MapReduce Programming

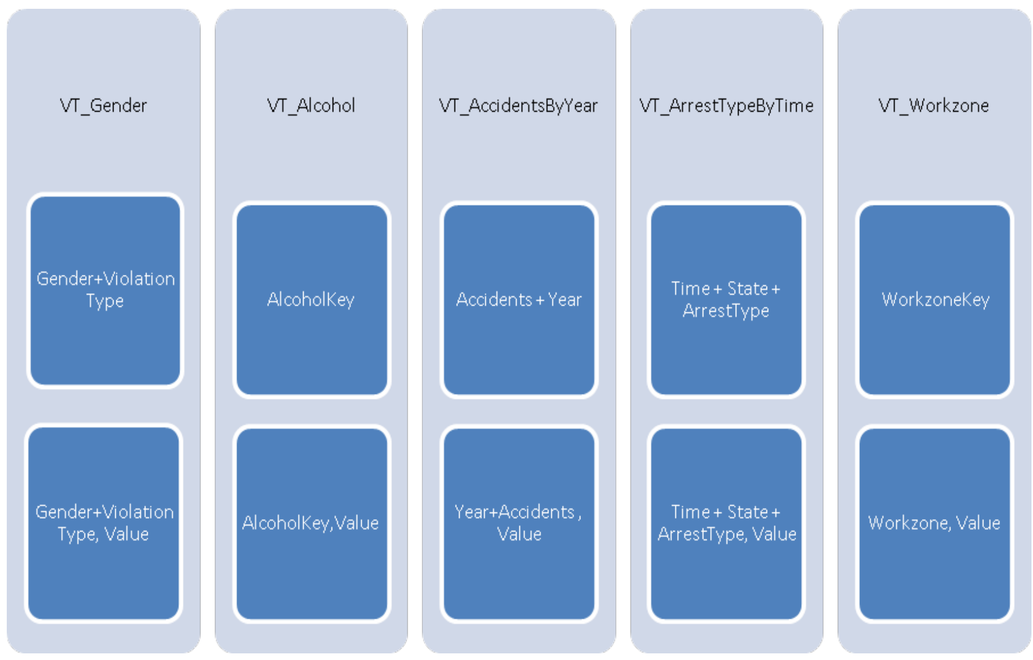

Fig 1

Results and Visualization

We have used Tableau tool for data analytics and visualization.
Traffic by Time, Date and Gender.

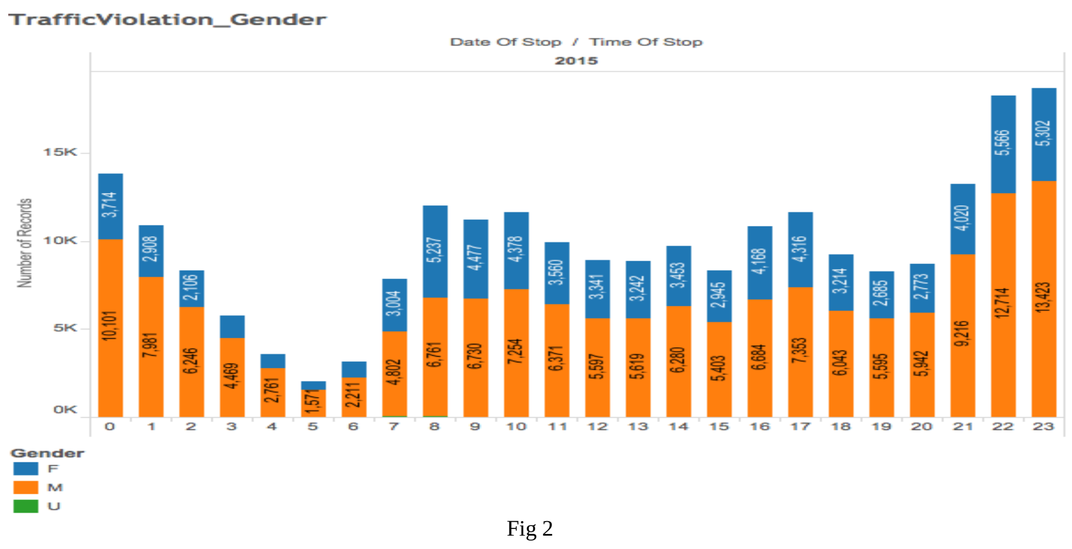

Fig 2

Traffic Violation Type by Gender

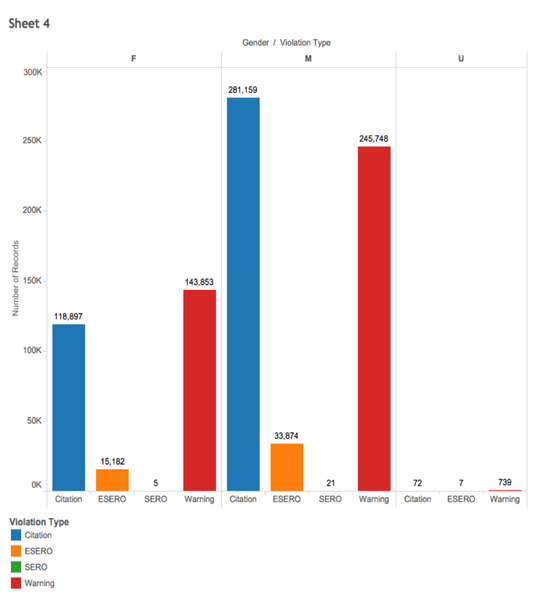

Fig 3

Traffic Violations / Accidents by Year

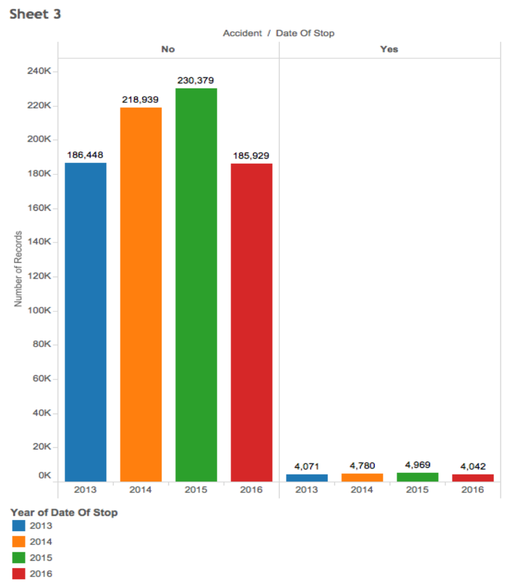

Fig 4

Conclusion

In this paper, data analytics for traffic violations in United States over 4 years has been presented. The proposed data modelling tools are MapReduce programming, Spark and Hive. This analysis is aimed at outlaying the various traffic violation patterns and assisting the police patrol department to intensify monitoring which particularly help save guide roads and enhance safety on roads. Observed patterns can also assist government in standardizing traffic rules and regulation, cater for road needs and decision making.